DOCUMENT RESUME

ED 337 499 TM 017 352

AUTHOR McNeil, Keith
TITLE Statistical Tests of Significance for the One Group Posttest Only Design.
PUB DATE Apr 91
NOTE 12p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Compensatory Education; Elementary Secondary Education; Evaluation Methods; *Pretests Posttests; *Program Evaluation; *Research Design; *Statistical Significance
IDENTIFIERS Education Consolidation Improvement Act Chapter 1; *Single Subject Designs; T Test

ABSTRACT

A research design is described for the situation in which a program, particularly a compensatory education program funded by Chapter 1 of the Hawkins Stafford Act of 1988, can be evaluated when there is no available comparison group and no pretest data. The design requires content specialists to identify which objectives on the posttest were included in the compensatory curriculum (C objectives) and which were included only in the regular curriculum (R objectives). Compensatory students should perform better on the C objectives to which they were exposed in both regular and compensatory curricula than on the R objectives. The analysis would be a simple t-test of the differences between two groups, the C items and the R items. The design is valuable because: (1) students serve as their own control group; (2) it is not necessary to identify a test that can measure pretest and posttest knowledge; and (3) it allows for identification of successful components of the Chapter 1 program. Two exhibits illustrate sample designs. Five figures and two tables present the analysis method and results from a 20-item test for 47 students in grades 1, 2, and 3 in 1987-88 and 34 students in 1988-89. An eight-item list of references is included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

KEITH McNEIL

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED337499

STATISTICAL TESTS OF SIGNIFICANCE FOR THE ONE GROUP POSTTEST ONLY DESIGN

Keith McNeil
New Mexico State University

TM 017352

BEST COPY AVAILABLE

Statistical Tests of Significance for the One Group Posttest Only Design

Keith McNeil
New Mexico State University

Problem: One of the purposes of evaluation is to foster usage of the evaluation results for program improvement. While most of my evaluation colleagues as well as myself sometimes blame administrators for not using evaluation results, the problem is often with the evaluation model, the type of feedback provided, or process used. Too often the evaluation model does not allow for specific recommendations. A good example is the evaluation models employed in the evaluation of Chapter 1.

Possible solutions: The avowed intent of the Chapter 1 evaluation models as developed by RMC and promulgated by the Technical Assistance Centers was to determine the effectiveness of the monies spent on Chapter 1 (Horst, Tallmadge, & Wood, 1975). The evaluation models were essentially of the "objectives-oriented" family, in that they accepted the objectives of the program and assumed that the program was implemented as described in the proposal that was funded. The evaluation models focused on the posttest performance, using the pretest or some proxy as an indicator of where the Chapter 1 students were supposed to be at the end of the year. In all fairness to the developers of the models, there was no intent of the models to identify which components of the program were not working or why those components were not working.

During the recent years of "search for excellence" and "school effectiveness", the Chapter 1 program office rightfully decided to push the Chapter 1 programs and the Technical Assistance Centers in the direction of "program improvement." As already indicated, though, the currently available models were not designed to assist in this endeavor. The current models do a good job of providing the "go -- no go" decision for the overall program, but provide no hint at all regarding the effectiveness of individual components. As a result, Chapter 1 Directors might look in other directions for evaluation tools. The accreditation model might be used by Directors, wherein they would invite experts to come into their program and provide an "expert opinion" as to the merits of the various components. The qualifications, experience, and biases of the "expert" may have a bearing on the evaluation results.

The "naturalist" models provide another option, wherein a relatively naive observer, using anthropological techniques, would spend time observing the project. As a result of being inundated in the project, the observer would then identify the pluses and minuses of the project from the point of view of the observer. A project will receive very different recommendations depending on who the naive person was and the degree of naivete. The

---

Paper Presented at the Annual Meeting of the
American Educational Research Association
April 3-7, 1991, Chicago, Illinois

TM017352

naturalistic method also usually requires an enormous amount of time and money.

The one group posttest only evaluation model has been identified as a relatively inexpensive and fruitful model (McNeil,1990a,1990b). The model can also identify which components of the curriculum are not being successful. The reasons for this lack of success would still need to be identified through other evaluation procedures, but the evaluation model has initiated a narrowing process.

Method: The one group posttest only design can be utilized to evaluate a compensatory program when there is no comparable comparison group and when pretest data do not exist (Ryan, 1980). The design requires content specialists to identify which objectives on the posttest were included in the compensatory curriculum (the C objectives), and which objectives were included only in the regular curriculum (the R objectives). Exhibit 1 provides a schematic representation of a 20 item test with the R and C designations. The compensatory students should perform better on those C objectives to which they were exposed in both the regular and the compensatory program (the double dosing effect), than on those R objectives that they were exposed to only in the regular curriculum.

Analysis: One could compare the percent correct on the items measuring the two groups of objectives. The analysis would be a simple t-test of the difference between two groups--one group being the C items and the other group being the R items, as indicated in Exhibit 1, producing a result as in Figure 1.

It is possible that the items measuring the one group of objectives are of different difficulty than the items that are measuring the other group of objectives. The solution to this potential dilemma is to statistically equate the difficulty of the items by covarying the inherent difficulty of the items. One could use the difficulty information from either: 1) the norming sample, 2) the non-compensatory students in the same school, 3) the results from the non-compensatory students in the same school in previous years, or 4) the results from one or more LEAs using the similar curriculum and similar in demographics. Since the difficulty information is used only as a covariate, the adequacy of the information is not too crucial. That is, the additional group is only providing information as to the difficulty of items on the posttest and the group is not being used as comparison group. The analysis would be a covariance analysis, covarying the difficulty of the items. The covariate is in the last column in Exhibit 1, and would produce a result as in Figure 2. Either of the above analyses can be performed on all of the items in the test or a subset.

When there is a desire to use the evaluation information for program improvement, one would want to analyze a specific subset of the items, such as:

. items can be reasonably grouped into curriculum units
. items can be grouped as to first semester or second semester
. items can be grouped into various taxonomic levels

An example of various taxonomic levels will be presented to illustrate the point. In Exhibit 2 the items in Exhibit 1 have been identified as (1) to which of three different classes following the taxonomy of Bloom (1956), and (2) to which semester they were supposed to be taught--first or second. The results in Figure 3 clearly show that the Chapter 1 students did better on the "knowledge" objectives that were in the Chapter 1 and Regular program than the objectives that were just in the Regular program. The results in Figure 3 are the kind of results that would be expected from the Chapter 1 students being double-dosed on the C objectives and only single-dosed on the R objectives.

Figure 4 indicates less success for the Chapter 1 students on "application" objectives. That is, Chapter 1 students are a little better on "application" objectives when they are double-dosed than when they get only the Regular dose of the "application" objectives.

Figure 5 indicates that the Chapter 1 students, and hence the Chapter 1 program, are not successful with "synthesis" objectives. Even though the Chapter 1 students received instruction on the "synthesis" objectives in both the Chapter 1 classroom and the regular classroom, they still did not perform any better on those double-dosed objectives than they did on the "synthesis" objectives which were only taught in the regular classroom.

Results such as those in Figure 3 would be expected. Taxpayers have paid extra money for the double-dosing and therefore rightfully expect higher performance on those objectives. Results such as those in Figure 4 are less exciting and might occur if teachers don't teach these "application" objectives as well as they should, or if Chapter 1 students don't learn these "application" objectives as well as they should. Possibly only a small amount of Chapter 1 time is spent on these "application" objectives, while the larger part of the Chapter 1 time is spent on the "knowledge" objectives in Figure 3.

Results such as those in Figure 5 are unacceptable and explanations such as those offered above need to be found. Perhaps Chapter 1 teachers were not provided enough inservice on how to teach "synthesis" objectives. Perhaps Chapter 1 teachers did not have enough time to include all of the material and purposefully left out the higher-order skill of "synthesis." Perhaps these Chapter 1 students did not receive enough support back in the regular classroom--perhaps they were led to believe that low achievers can not be successful on higher-order skills such as "synthesis."

The one group posttest only model has thus identified program problems in the area of Synthesis. The specific reason for lack of success would have to be identified through additional evaluation procedures, such as:

. Evaluation of staff development to determine if inservice emphasized Synthesis objectives as much as other aspects of the Chapter 1 curriculum, or

. Observation of Chapter 1 teachers to determine if the lesson plans allowed for enough time to teach "synthesis", or

. Observation or questioning of Regular teachers to determine if Chapter 1 students received equal encouragement on all the objectives in the regular classroom.

Special concerns: The design rests heavily on the accuracy of the curriculum specialists being able to identify those objectives that were included in the two curricula. The task can be made a little easier by using a criterion-referenced test that has been designed to measure the regular curriculum. In such a case, the content people only have to identify those objectives that are in the compensatory curriculum.

In most school systems there is the additional assumption that the teachers actually taught the curriculum (and that the students listened to and learned from the curriculum). The extent to which these assumptions are tenable causes problems for all evaluation models, but only reduces the likelihood of obtaining significant results in favor of the compensatory program in the one group posttest only design.

Potential problems: Since this is a new design, one might wonder about whether or not there might be some problems in implementing the design. The author successfully implemented this design in two successive years in a Chapter 1 program in Dallas (McNeil, Berry, & Metze, 1988: McNeil,Jones, Berry, Edoghotu,& Kane). Tables 1 and 2 present the results from that application. Several potential problems, though, might be considered.

Calculations. As with any new evaluation model, ease in implementation is a reasonable concern. Analysis I in Figure 1 is a straight-forward computation of the difference between two means. Analysis II in Figures 2-5 requires an evaluator who understands covariance. For those who understand this concept, the interpretive value of this analysis far outweighs the additional calculation burden. Existing computer packages such as SAS and SPSS can easily perform the calculations.

Aggregation of data. State and Federal evaluators want the data to be collapsible across LEAs. If the data are transformed to logits, a fairly straight-forward procedure, one should be able to aggregate the results. On the other hand, one could argue that evaluation for program improvement should be oriented to the project, and not to the aggregation needs of the Federal government. The Dallas application resolved this problem by using the one group posttest only design for local use and the traditional models for reporting to State and Federal agencies.

Interpretation of results. The interpretation of results will have to rely on usage over time, as did the NCE metric when it was first introduced. It should be clear by now that the objective level interpretations provide insights into curriculum, inservice, and teaching modifications that are not available with the current Chapter 1 evaluation models.

Determination of which curriculum items are in. This determination probably needs to be made by content specialists, rather than by evaluators. The task can be difficult and time consuming. On the other hand, one might argue that the content specialists should know both the regular and compensatory curricula well enough so that the task would not be that difficult, as was the case in the Dallas application. In addition, such determinations are usually made when an LEA makes a test adoption decision. (One added benefit of this design is that the test adoption decision is less crucial for the compensatory program.

Those items that are not in an LEA's curriculum or in the compensatory curriculum can be omitted from the analysis, which is not possible in the RMC evaluation models.)

Teacher implementation of curriculum. If the Chapter 1 teachers do not implement the Chapter 1 program as expected, then the analysis will wrongly accuse the Chapter 1 program of being not effective. Observation of Chapter 1 teachers could avoid this conclusion.

Only low difficulty items in the curriculum. A Chapter 1 curriculum might focus on low-level objectives, but most tests are designed such that each objective is measured by items across the range of difficulty. If indeed the Chapter 1 curriculum is measured only by items of low difficulty, then analysis I will lead to an incorrect conclusion, but analysis II will still be applicable.

Testing out of level. Many compensatory students take a lower level test, as recommended by the developers of the Chapter 1 evaluation models (Roberts,1981). Since the same kind of curriculum fit determinations can be made with an out of level test as with an on level test, testing out of level would not cause a problem with the new evaluation model.

Summary: An evaluator may on occasion be confronted with the need to produce an evaluation of a compensatory program when there is no available comparison group and when no pretest data is available. The design discussed in this paper provides a tool for obtaining an evaluation under such constraining circumstances, without sacrificing any evaluation principals.

The design is particularly valuable for three reasons. First, few, if any, evaluators ever find a perfect comparison group in the real world. In this design, the students serve as their control. Second, if program gains are evaluated over a school year, which they usually are, it may be inappropriate to use the same test for both pretest and posttest. It may be very difficult to identify a test which adequately measures the objectives desired at the posttest and which can be administered at pretest. Finally, and most importantly for the this paper, the design allows for the identification of which components of the Chapter 1 program are successful and which are not so successful, providing guidance for program improvement decisions.

NOTE: I would like to thank Joe Ryan for initially discussing this design, and Napoleon Mitchell, Gail Smith, Wayne Murray, William Denton, George Powell, James English, and David Vines for forcing me to have a better conceptualization of the design. I especially want to thank Barbara Mathews, Jane Seibert, and Rosie Ramirez for identifying the items and helping me chart the unknown.

Item #	In Regular Curriculum	In Chapter 1 Curriculum	Item Designation	Posttest Percent Correct	Inherent Difficulty
1	Y	Y	C	.40	.40
2	Y	Y	C	.78	.68
3	Y	Y	C	.80	.85
4	Y	N	R	.30	.40
5	Y	N	R	.68	.78
6	Y	N	R	.10	.20
7	N	N	OMIT	.20	.40
8	N	Y	OMIT	.50	.78
.					
.					
.					
20	Y	Y	C	.20	.15

Exhibit 1. Sample design.

Item #	Taxonomic Level	Semester Planned	Item Designation	Posttest Percent Correct	Inherent Difficulty
1	knowledge	first	C	.40	.40
2	application	first	C	.78	.68
3	synthesis	second	C	.80	.85
4	knowledge	first	R	.30	.40
5	application	second	R	.68	.78
6	synthesis	first	R	.10	.20
7	application	first	OMIT	.20	.40
8	synthesis	second	OMIT	.50	.78
.					
.					
.					
20	application	first	C	.20	.15

Exhibit 2. Sample design, with program improvement application.

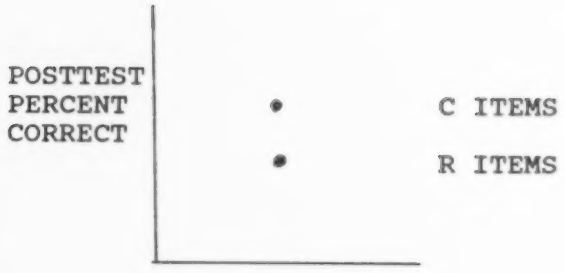

Figure 1. Schematic results from analysis I, two group means.

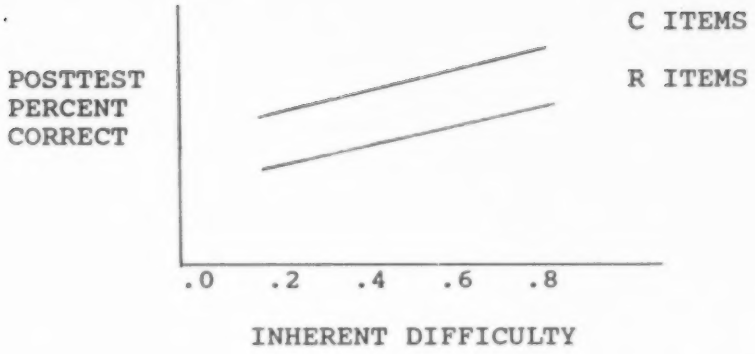

Figure 2. Schematic results from analysis II, inherent difficulty as covariate.

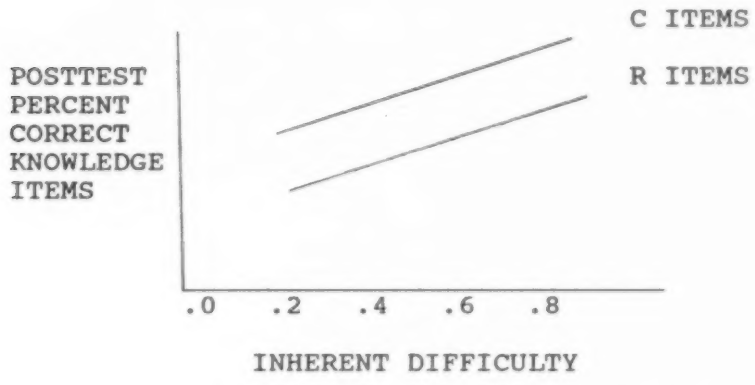

Figure 3. Schematic results from analysis II, on Knowledge items.

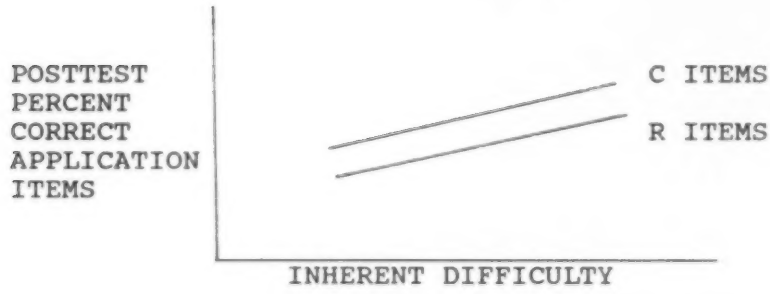

Figure 4. Schematic results from analysis II, on Application items.

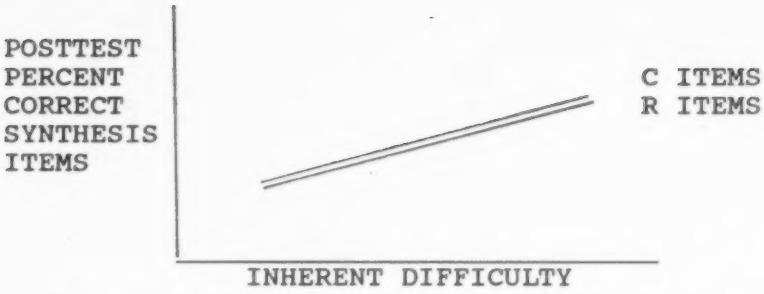

Figure 5. Schematic results from analysis II, on Synthesis items.

Table 1. Percent Correct on STEELS Language Arts Items Included and Not Included in the A Priori Curriculum, 1987-88.

	Items Included in A Priori		Items Not Included in A Priori		
Grade	Percent Correct	N	Percent Correct	N	Probability of Difference
1	70.1	13	66.9	20	.009
2	73.3	18	71.9	23	.205
3	66.9	16	65.1	21	.124
All	70.1	47	68.2	64	.002

Note. Items were adjusted for overall difficulty.

Table 2. Percent Correct on STEELS Language Arts Items Included and Not Included in the A Priori Curriculum, 1988-89.

	Items Included in A Priori		Items Not Included in A Priori		
Grade	Percent Correct	N	Percent Correct	N	Probability of Difference
2	70.0	18	72.4	23	.72
3	70.8	16	64.5	21	.04
All	70.4	34	68.3	44	.12

Note. Items were adjusted for overall difficulty.

REFERENCES

Bloom, B. S. (1956) Taxonomy of Educational Objectives. New York: David McKay Company.

Horst, D.P., Tallmadge, G.K., and Wood, C.T. A practical guide for measuring project impact on student achievement. Washington, D.C.: U. S. Government Printing Office, 1975 (Stock No. 107-080-01460).

McNeil, K. (1990, January). The one group posttest only evaluation model. Paper presented at the meeting of the Southwest Educational Research Association. Austin, TX.

McNeil, K. (1990, October). Use of the new one group posttest only design. Paper presented at the meeting of the Midwestern Educational Research Association. Chicago, IL.

McNeil, K., Berry, R., and Metze, B. (1988, July). Chapter 1 Basic Skills Final Evaluation Report, 1987-88 (REIS88-001-7). Dallas, Texas: Dallas Independent School District, Department of Research, Evaluation and Information Systems.

McNeil, K., Jones, K., Berry, R., Edoghotu, F., and Kane, R. (1989, July). Evaluation of the 1988-89 Chapter 1 Instructional Program (REIS89-001-5). Dallas, Texas: Dallas Independent School District, Department of Research, Evaluation and Information Systems.

Roberts, A.O.H. Out-of-level testing. In Evaluator's References: Title 1 Evaluation and Reporting System Vol 2. Washington, D.C.: U.S. Government Printing Office, 1981 (Stock No. 728-190-1792).

Ryan, J. Personal communication, 1980.